

# 大规模知识图谱技术.pdf

## 知识图谱实例

- 谷歌 知识图谱 (Google Knowledge Graph)
- 百度 “知心”
- 搜狗 “知立方”

## 知识图谱的构建

- 知识图谱的数据来源

mainly for extraction graphs(抽取图谱)

- 百科类数据
  - 质量较高
  - 更新速度慢
  - 提取实体、属性、关系
- 结构化数据
  - 方法：构建面向站点的包装器
  - 质量较高
  - 更新速度慢
  - 提取实体、属性、关系，属性部分加强
- 搜索日志
  - 质量较差
  - 提高图谱覆盖率
  - 提取属性
- 从抽取图谱到知识图谱
  - 实体对齐 (Object Alignment)
    - 目的：发现具有不同标识却代表真实世界中同一对象的那些实体，并将这些实体归并为一个具有全局唯一标识的实体对象，然后添加到知识图谱中
    - 主要方法：聚类
      - 相似度度量规则
        1. 具有相同描述的实体可能代表同一实体（字符相似）
        2. 具有相同属性 - 值的实体可能代表相同对象（属性相似）
        3. 具有相同邻居的实体可能指向同一个对象（结构相似）
      - 准确率无法保证，需人工审核
    - 知识图谱模式构建
      - 本体
        - 概念
        - 概念层次
        - 属性
        - 属性值类型
        - 关系
        - 关系定义域概念集
        - 关系值域概念集
      - 图谱模式
        - 领域 (domain)
        - 类别 (type)
        - 主题 (topic, 即 实体)
      - 方法
        - 自底向上

有利于抽取新的实例，可保证抽取质量
        - 自顶向下

能发现新的模式
    - 推理

推理 ( reasoning 或 inference ) 被广泛用于发现隐含知识，其功能通过可扩展的规则引擎来完成

      - 规则
        - 针对属性
        - 针对关系

- 实体重要性排序
- 相关实体挖掘

“其他人还搜了”

- 知识图谱的更新和维护
  - 知识图谱模式的更新
    - 目前定义类别数约为 103~104 量级
    - 由专业人员进行决策和命名新类别
  - 结构化站点包装器的维护
    - 变化量超过事先设定的阈值且抽取结果与原先标注的答案差别较大，则表明现有的站点包装器失效了
  - 知识图谱的更新频率
    - 规模和更新频度：数据层>>模式层
  - 众包反馈机制

#### 知识图谱在搜索中的应用

- 查询理解
  1. 选择性显示知识卡片
  2. 选择性显示属性
- 问题回答

#### 总结

1. 目前知识图谱的发展还处于初期阶段
2. 人工干预仍起重要作用
3. 结构化数据在知识图谱的构建中起到决定性作用
4. 各大搜索引擎公司为了保证知识图谱的质量多半采用成熟的算法
5. 搜索引擎公司展示知识卡片时比较谨慎
6. 更复杂的自然语言查询将崭露头角（如谷歌的蜂鸟算法）

---

知识图谱：旨在描述真实世界中存在的各种实体或概念

知识卡片：用户查询中所包含的实体或返回的答案提供的详细结构化摘要，是特定查询的知识图谱

实体：全局唯一确定的标识符

属性值对：( attribute-value pair, AVP ) 实体特性

关系：实体间关联

---

#### 多学科结合

- 知识库
- 自然语言处理
- 机器学习
- 数据挖掘